

Robust Morphometric Analysis based on Landmarks. Applications

A. García-Pérez and M.Y. Cabrero-Ortega

Abstract Procrustes Analysis is a Morphometric method based on Configurations of Landmarks that estimates the superimposition parameters by least-squares; for this reason, the procedure is very sensitive to outliers. In the first part of the paper we robustify this technique to classify individuals from a descriptive point of view. In the literature there are also classical results, based on the normality of the observations, to test whether there are significant differences between individuals. In the second part of the paper we determine a Von Mises plus Saddlepoint approximation for the tail probability of the Procrustes Statistic when the observations come from a model close to the normal. We conclude the paper with some applications using the Geographical Information System QGIS.

Keywords: Robustness; Morphometrics; Von Mises expansion; Saddlepoint approximations; Geographical Information System QGIS

1 Introduction

This paper is about a robust classification problem of n individuals based on their shapes, i.e., using their geometric information. The usual (classical or robust) methods based on a Multivariate Analysis can not extract all the geometric information from the individuals. For this reason, in recent years morphometrics methods based on Configurations of landmarks have

A. García-Pérez (✉)

Departamento de Estadística, I.O. y C.N., Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 9, 28040-Madrid, Spain
e-mail: agar-per@ccia.uned.es

M.Y. Cabrero-Ortega

C.A. Universidad Nacional de Educación a Distancia (UNED)-Madrid, Spain
e-mail: ycabrero@madrid.uned.es

been developed. A landmark is a peculiar point whose position is common in all the individuals to classify. For instance, when we classify skulls, the landmarks could be the center of the supraorbital arch, the chin, etc.; or, if we classify projectile points found in an archaeological site, the landmarks could be the ends of the points.

In all the cases, the mathematical (geometric) information that we obtain from individuals is the k coordinates of their p landmarks, $l_i = (c_{i1}, \dots, c_{ik})$, $i = 1, \dots, p$.

The matrix of landmarks coordinates is called a Configuration. For each individual with p landmarks of dimension k (2 or 3) we shall have a collection of landmark coordinates expressed in $p \times k$ matrix as

$$M = \begin{pmatrix} c_{11} & \cdots & c_{1k} \\ \vdots & \vdots & \vdots \\ c_{p1} & \cdots & c_{pk} \end{pmatrix}$$

2 Classical Morphometric Analysis from a Descriptive Point of View

As we have mentioned before, we shall use the shape of the individuals in their classification. Shape is a property of an object invariant under scaling, rotation and translation; otherwise, for instance, an object and itself with double size could be classified into two different groups.

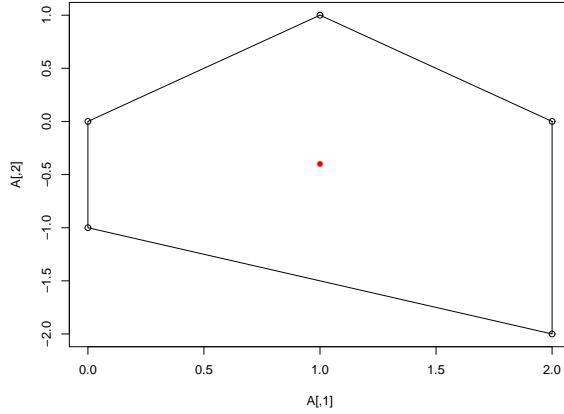


Fig. 1 Polygon representation of a Configuration with 5 landmarks

There are many morphometric methods; see for instance [1] or [3]. In this paper we shall consider Superimposition Methods; namely, Procrustes Analysis, obtaining the Procrustes coordinates with it, adapting the Configurations to a common (local) reference system and matching them at the common center. For these reasons, a Local Coordinate Reference System is needed and a Geographical Information System will be very useful.

A common graphical representation of a Configuration is a scatter plot of its landmarks coordinates. Joining them with segments we obtaining a polygon as, for instance, in Fig. 1, where the landmarks coordinates are the vertices of the polygon.

As we have said above, to classify individuals we have first to remove the effect of Size (scale), Location (translation) and Orientation (rotation) to standardize them and match them in a common center (the centroid of the polygon) in order to make them comparable.

To apply the Procrustes superimposition method we have to estimate by least-squares the superimposition parameters α , β and Γ (scale, translation and rotation) in order to minimize the full Procrustes distance d_F between Configurations M_1 and M_2 , i.e.,

$$\begin{aligned} \min d_F(M_1, M_2) &= \min \|M_2 - \alpha M_1 \Gamma - 1_p \beta'\| = \\ &= \sqrt{\text{trace}[(M_2 - \alpha M_1 \Gamma - 1_p \beta')'(M_2 - \alpha M_1 \Gamma - 1_p \beta')]} \end{aligned}$$

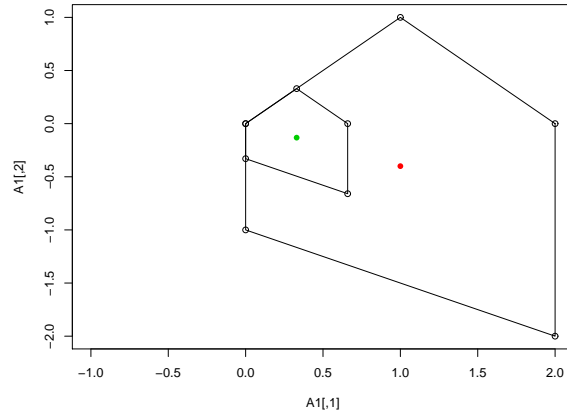


Fig. 2 Polygons representing a Configuration with 5 landmarks: the original one with red centroid and the scaled to Centroid-mean Size equal to 1 with green centroid

where α is a scalar representing the Size, β is a vector of k values corresponding to a Location parameter formed by the centroid coordinates, 1_p is a column vector of dimension $p \times 1$ and Γ a $k \times k$ square rotation matrix.

The idea that we pursue with this transformation is to match both Configurations, i.e., a superimposition of M_1 onto M_2 .

2.1 Removing the Size Effect

The first step we must take in Procrustes Analysis to standardize Configurations is to remove the *Size* effect. If, as usual, we consider as center the *centroid-mean* of dimension k (sample mean by columns) defined by

$$M_c = (M_{c1}, \dots, M_{ck}) = \left(\frac{1}{p} \sum_{i=1}^p c_{i1}, \dots, \frac{1}{p} \sum_{i=1}^p c_{ik} \right)$$

and easily computed with R as ([10])

```
> apply(M,2,mean)
```

the *Centroid-mean Size* is defined as

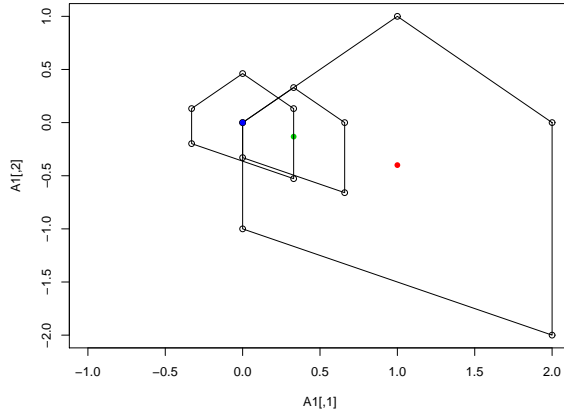


Fig. 3 Polygons representing a Configuration with 5 landmarks: the original one with red centroid, the scaled to Centroid-mean Size equal to 1, with green centroid, and the centered with respect location and translation with blue centroid

$$CS = \sqrt{\sum_{i=1}^p d_E^2(l_i, M_c)} = \sqrt{\sum_{i=1}^p \sum_{j=1}^k (c_{ij} - M_{cj})^2} = \sqrt{\sum_{j=1}^k p \cdot \text{Var}(c_{.j})}$$

being $d_E^2(l_i, M_c)$ the square of the Euclidean distance between the i th landmark l_i and the centroid-mean M_c . Hence, the Centroid-mean Size depends on the sample variance and so, it will be very sensitive to outliers. This size can be computed as

```
> sqrt(sum(apply(M,2,var)*(p-1)))
```

The coordinates of a scaled Configuration are now calculated dividing the original coordinates by CS

$$M_{cs} = \frac{M}{CS}$$

In Fig. 2 we see the previous Configuration (with red centroid) and the scaled to Centroid-mean Size equal to 1 (the Configuration with green centroid).

2.2 Removing Location by Translation

We remove the Location effect translating the Configuration matrix so that its centroid is the new origin. We do this with the R sentence

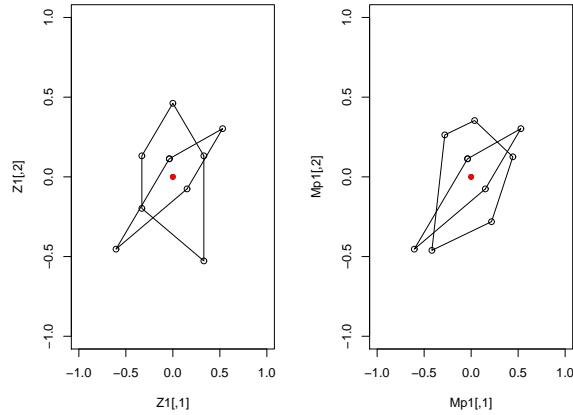


Fig. 4 Polygon rotated

```
> scale(M,scale=F)
```

In Fig. 3 we have the previous Configurations and the centered one (with blue centroid).

2.3 Removing Orientation by Rotation

After the effect of Size and Location have been removed, we estimate (by least-squares) the rotation matrix Γ minimizing the distance between Configurations M_1 and M_2 , i.e.,

$$\min_{\Gamma} \|M_2 - M_1 \Gamma\| = \min_{\Gamma} \sqrt{\text{trace}((M_2 - M_1 \Gamma)'(M_2 - M_1 \Gamma))}$$

where Γ is a $k \times k$ square rotation matrix, a matrix that must be determined in order to maximize the correlation between the two sets of landmarks, i.e., to minimize the distance between landmarks. More precisely:

If M_1 and M_2 are two Configurations and X_1 and X_2 the corresponding centered Configurations scaled to unit Centroid-mean Size, the (full) Procrustes distance is defined as

$$d_F(M_1, M_2) = \sqrt{\text{trace}(X_2 - \beta X_1 \Gamma)'(X_2 - \beta X_1 \Gamma)} =$$

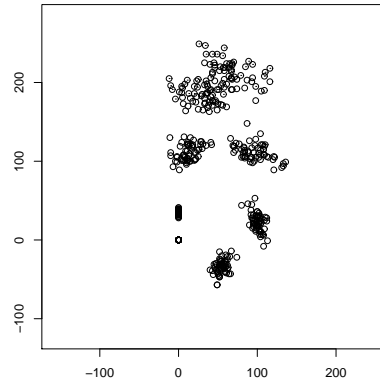


Fig. 5 Scatter plot of landmarks of Example 1

$$= \sqrt{1 - \left(\sum_{j=1}^k \lambda_j \right)^2}$$

where λ_j are the diagonal elements of matrix D in the singular-value decomposition of $X_2'X_1$:

$$X_2'X_1 = UDV'$$

But in fact, the previous problem is a known mathematical issue: If we have the previous singular-value decomposition of $X_2'X_1$, the rotation matrix we are looking for is

$$\Gamma = VU'$$

In Fig. 4 we have two Configurations, before and after rotated one of them, according to the previous method.

2.4 More than two Configurations (Generalized Procrustes Analysis)

In the previous sections we have done, in three steps, what is called a classical *Partial Procrustes Analysis* because we have compared, from a descriptive point of view, two Configurations.

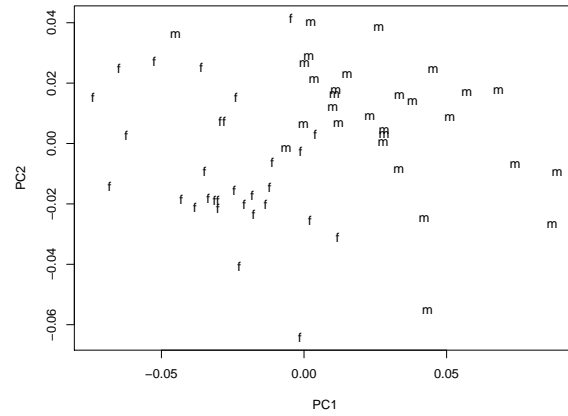


Fig. 6 Scores of the Principal Component analysis

If we have more than two Configurations we have to do what is called a *Generalized Procrustes Analysis* in three steps:

1. All the Configurations must be standardized, i.e., their centroids matched at a common origin and scaled to unit size.
2. We have to define a consensus or average Configuration as a reference, called *Mean Shape* because, in fact, it is the sample mean of all the Configurations, by vertices, i.e., the mean of the vertices (i, j) (homologous coordinates) for the n Configurations. The Mean Shape Configuration can be computed as

```
> apply(M,c(1,2),mean)
```

3. Finally, we perform a (partial) Procrustes superimposition between the Mean Shape and each Configuration.

2.5 Configuration Projection onto the Tangent Space

The shape space defined by the previous Procrustes superimposition method is non-Euclidean and corresponds to a curved surface. This means that the distance between two landmarks is not the length of the segment joining them and hence, we cannot apply traditional statistics to the Procrustes coordinates of the n individuals.

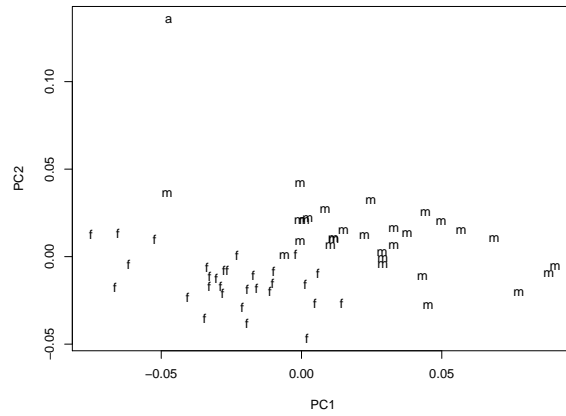


Fig. 7 Scores of the Principal Component analysis with the outlier a

From a convenient point of view, the n individuals are *aligned* in the $n \times kp$ matrix X and then projected onto the (Kendall) tangent space, where the vectorized Mean Shape x_m (i.e., a vector of dimension $kp \times 1$) is the contact point between spaces. The projected (tangent) coordinates are obtained in a matrix X^* as

$$X^* = X(I_{kp} - x'_m x_m)$$

where I_{kp} is the $kp \times kp$ identity matrix.

Then, we can apply the usual statistical techniques to these projected coordinates, for example, classifying the resulting observations with the scores of their Principal Components.

Example 1. In paper [9], 59 gorilla skulls were considered. We know, in the example but not in a real case, that 30 of them are female and 29 male. In their paper, 8 landmarks were considered. If we represent these data in a scatter plot we obtain Fig. 5 where no apparent classification between males and females is observed.

If we make the four previous steps of the Generalized Procrustes analysis and conclude with a Principal Component analysis of the scores, we obtain Fig. 6 where we cannot appreciate the two groups very clearly although the vertical bar at $PC1=0$ is the usual classification rule taken for this example.

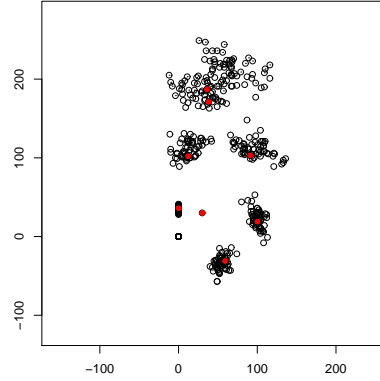


Fig. 8 Scatter plot of the 59 gorillas plus the outlier **a** in red

3 Robust Morphometric Analysis from a Descriptive Point of View

Let us consider again the data of Example 1 plus a Configuration, that we call **a**, for which we replace the coordinates of the third landmark as in the following diagram:

	[,1]	[,2]					[,1]	[,2]	
[1,]	36	187					[1,]	36	187
[2,]	59	-31					[2,]	59	-31
[3,]	0	0	----->				[3,]	30	30
[4,]	0	36		a =			[4,]	0	36
[5,]	12	102					[5,]	12	102
[6,]	38	171					[6,]	38	171
[7,]	91	103					[7,]	91	103
[8,]	100	19					[8,]	100	19

If we give the four previous steps to perform a Generalized Procrustes analysis, we obtain the classification given in Fig. 7 where all the individuals are in one group except outlier **a**.

But in Fig. 8 we see that **a** is in the bulk of the data and also the mean shape in Fig. 9. Hence, no apparent solution is clear.

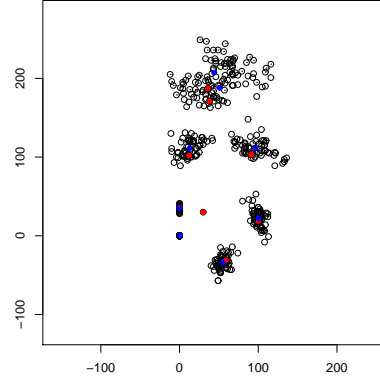


Fig. 9 Scatter plot of the 59 gorillas plus the outlier **a** in red and the mean shape in blue

3.1 Removing the Size Effect in a Robust Way

We propose, instead of using the centroid-mean

$$M_c = (M_{c1}, \dots, M_{ck}) = \left(\frac{1}{p} \sum_{i=1}^p c_{i1}, \dots, \frac{1}{p} \sum_{i=1}^p c_{ik} \right)$$

as before, that essentially is a sample mean computed with

```
> apply(M,2,mean)
```

to use the median (or the trimmed-mean) by columns with the following two R sentences,

```
> apply(M,2,median)
```

```
> apply(M,2,mean,trim = .2)
```

obtaining in this way a more robust centroid. Now, instead of considering the Centroid-mean Size CS that, as we saw before, is essentially the variance

$$CS = \sqrt{\sum_{i=1}^p d_E^2(l_i, M_c)} = \sqrt{\sum_{i=1}^p \sum_{j=1}^k (c_{ij} - M_{cj})^2} = \sqrt{\sum_{j=1}^k p \cdot Var(c_{.j})}$$

computed with

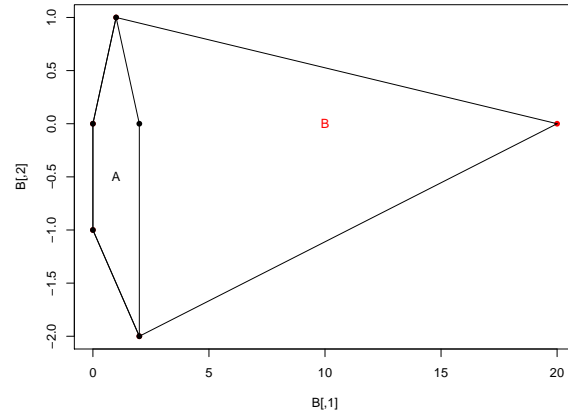


Fig. 10 Two configurations with a very different size

```
> sqrt(sum(apply(M,2,var)*(p-1)))
```

we propose to use the Median Absolute Deviation MAD , defined as

$$MAD = 1.4826 M_e \{|X_i - M_e(X_i)|\}$$

obtaining with it what we call the *Centroid-median Size*

$$MS = \sum_{j=1}^k MAD(c_j)$$

computed as

```
> sum(apply(M,2,mad))
```

and that satisfies the Size invariance property $MS(aM) = aMS(M)$ for any positive scalar a .

In this way we obtain a robust size measure. For instance, considering the following two configurations A and B

<pre>> A</pre> <table style="border: none; margin-left: 40px;"> <tr> <td style="padding: 0 10px;">[,1]</td> <td style="padding: 0 10px;">[,2]</td> </tr> <tr> <td style="padding: 0 10px;">[1,]</td> <td style="padding: 0 10px;">2 0</td> </tr> </table>	[,1]	[,2]	[1,]	2 0	<pre>> B</pre> <table style="border: none; margin-left: 40px;"> <tr> <td style="padding: 0 10px;">[,1]</td> <td style="padding: 0 10px;">[,2]</td> </tr> <tr> <td style="padding: 0 10px;">[1,]</td> <td style="padding: 0 10px;">20 0</td> </tr> </table>	[,1]	[,2]	[1,]	20 0
[,1]	[,2]								
[1,]	2 0								
[,1]	[,2]								
[1,]	20 0								

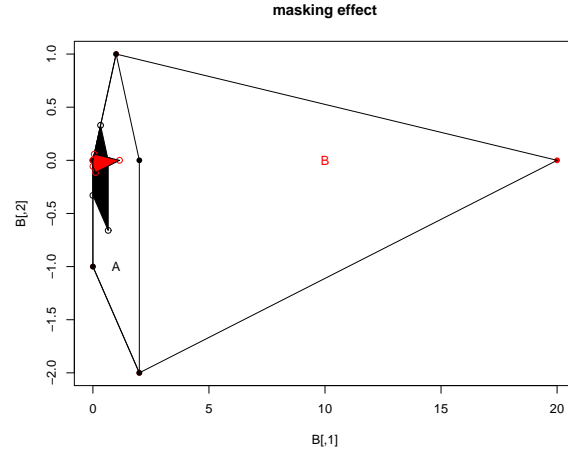


Fig. 11 Two configurations with a very different size after being standardized with the centroid-mean size

[2,]	1	1	[2,]	1	1
[3,]	0	0	[3,]	0	0
[4,]	0	-1	[4,]	0	-1
[5,]	2	-2	[5,]	2	-2

that differ in just a wrong digit in the first landmark of Configuration B, the classical Centroid-mean Size is very sensitive:

```
> sqrt(sum(apply(A,2,var)*4))
[1] 3.03315
> sqrt(sum(apply(B,2,var)*4))
[1] 17.44706
```

but not the the Centroid-median Size:

```
> sum(apply(A,2,mad))
[1] 2.9652
> sum(apply(B,2,mad))
[1] 2.9652
```

And what is more importante, this new size measure keeps the relative size of the Configurations avoiding a possible masking effect. For instance, in Fig. 10, if we divide both Configurations by the classical Centroid-mean

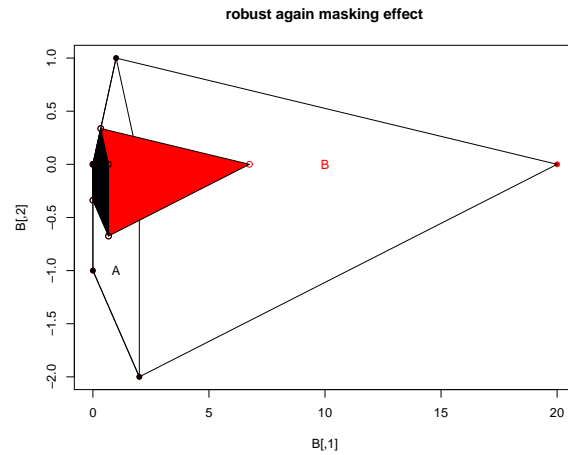


Fig. 12 Two configurations with a very different size after being standardized with the centroid-median size

Size we obtain Fig. 11 and they would probably be classified in the same final group. Nevertheless, standardizing them with the new robust Centroid-median Size, we see in Fig. 12 that the differences in size between them, remain.

Hence, instead of dividing the coordinates of the Configuration by the classical Centroid-mean Size CS , we propose to divide the configuration M (the coordinates) by the robust Centroid-median Size to distinguish between individuals in a better way, avoiding a possible masking effect,

$$M_{rs} = \frac{M}{MS}$$

3.2 Removing Location in a Robust Way

In the same way as we have removed the location effect in a classical way, translating the Configuration matrix so that its centroid-mean was the new origin, with the sentence

```
scale(M,scale=F) = scale(M,scale=F,center=apply(M,2,mean))
```

subtracting the mean of each column to the whole column, in the robust version we subtract the median with the sentence

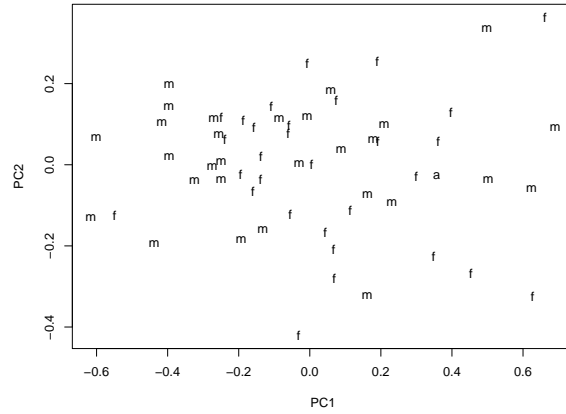


Fig. 13 Classification of gorillas with the median as *mean shape*

```
> scale(M,scale=F,center=apply(M,2,median))
```

being the centroid-median the new origin.

After robustifying with respect scale and location we keep the classical rotation matrix for the robust coordinates. These three steps are in our new R function `rpgpa1` that can be obtained on request from the authors. We conclude the process with the same coordinates projection formula than before.

3.3 More than two Configurations

If there are more than two Configurations, the key point in the robustification process is the selection of a robust *mean shape*, that in the classical Morphometric analysis is the sample mean of the Configuration coordinates. In our robust version we propose to choose as *mean shape* the median of the Configuration coordinates, obtaining Fig. 13 for the gorillas example (after doing a classical Principal Component analysis of the scores).

Considering the 0'2-trimmed mean as *mean shape* we obtain Fig. 14. Finally, considering the 0'1-trimmed mean as *mean shape* we obtain Fig. 15.

These three options are in new R function `rpgpa2` that can be composed with `rpgpa1`.

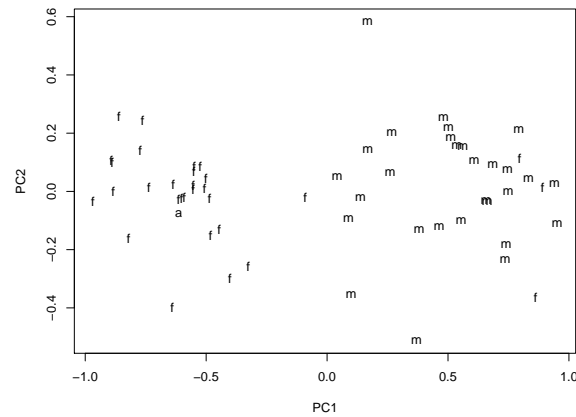


Fig. 14 Classification of gorillas with the 0'2-trimmed mean as *mean shape*

4 Classical Morphometric Analysis from an Inferential Point of View

Instead of considering a descriptive morphometric analysis it is more interesting to test if there are significant differences between two Configurations. From a classical point of view, we have the following result in [7] and [11]: If X_1 and X_2 are two scaled and centered Configurations with $p \times k$ landmarks, the *Residual Distance* between Configurations X_1 and X_2 is defined as

$$\|X_2 - X_1\|^2 = \text{trace} [(X_2 - X_1)'(X_2 - X_1)].$$

As saw in the previous sections, the $k \times k$ square rotation matrix Γ is determined such that the Procrustes distance between these two Configurations X_1 and X_2 (i.e., between landmarks) is minimal

$$\min_{\Gamma} \|X_2 - X_1 \Gamma\|^2 = \min_{\Gamma} \text{trace} [(X_2 - X_1 \Gamma)'(X_2 - X_1 \Gamma)].$$

This minimum (i.e., after matching, i.e., after translation, rotation and scaling) that we obtain is called the *Procrustes statistic*:

$$G(X_1, X_2) = \min_{\Gamma} \|X_2 - X_1 \Gamma\|^2.$$

Under the null hypothesis H_0 that there is no systematic difference between Configurations X_1 and X_2 , i.e., they belong to the same group, or more precisely, that they are of the form

$$X_2 = X_1 + \eta e$$

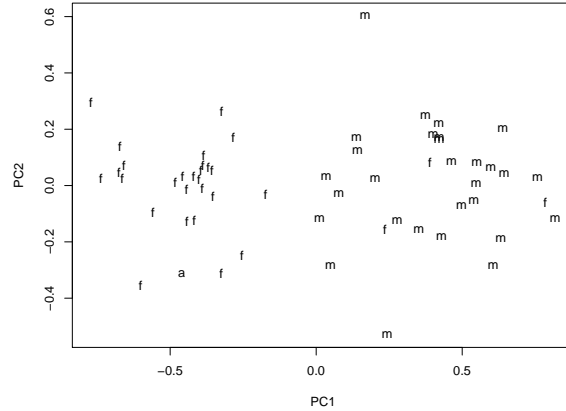


Fig. 15 Classification of gorillas with the 0'1-trimmed mean as *mean shape*

where all the elements of Configuration e are i.i.d. $N(0, 1)$, then

$$G(X_1, X_2) \approx \eta^2 \chi_g^2$$

i.e., $G_s(X_1, X_2) = G(X_1, X_2)/\eta^2 \approx \chi_g^2$, where $g = kp - k(k+1)/2 - 1$ obtaining so, a way to compute the tail probabilities (p-values) for testing H_0 . It must be $p > (k+1)/2 + 1/k$ and obviously an integer.

5 Robust Morphometric Analysis from an Inferential Point of View

The standard normality of the landmarks is a very difficult assumption to assume and check. For this reason we shall use robust methods for testing H_0 assuming that the $p \times k$ landmarks of e follow, not a standard normal distribution but a contaminated normal model:

$$\frac{X_2 - X_1}{\eta} \rightsquigarrow (1 - \epsilon)N(0, 1) + \epsilon N(0, \nu)$$

In this section we are going to compute the tail probabilities (p-values), assuming this contaminated model, using a VOM+SAD approximation.

We use this scale contaminated normal mixture model because the Configurations are matched at the common centroid that is the new origin and equal to 0, being the contamination in the scale the source of contamination in the observations.

5.1 Von Mises Approximations for the p-value of the Procrustes Statistic

In order to test the null hypothesis H_0 that there is no systematic difference between the standardized Configurations X_1 and X_2 , using the Procrustes statistic $G_s(X_1, X_2)$ that follow a χ_g^2 distribution under a normal model, we have the following result.

Proposition 1. *Let $G_s(X_1, X_2)$ be the Procrustes statistic, that follows a χ_g^2 distribution when the underlying model is a normal distribution, $\Phi_{\mu, \sigma}$. If the previous null hypothesis H_0 holds, the von Mises (VOM) approximation for the functional tail probability (if F is close to the normal $\Phi_{\mu, \sigma}$) is*

$$P_F\{G_s(X_1, X_2) > t\} \simeq g \int_{-\infty}^{\infty} P\{\chi_{g-1}^2 > t - (\frac{x-\mu}{\sigma})^2\} dF(x) - (g-1)P\{\chi_g^2 > t\}.$$

Proof. The von Mises (VOM) approximation for the functional tail probability is (if F is close to the normal $\Phi_{\mu,\sigma}$)

$$p_g^F = P_F\{G_s(X_1, X_2) > t\} \simeq p_g^\Phi + \int \text{TAIF}(x; t; \chi_g^2, \Phi_{\mu,\sigma}) dF(x) \quad (1)$$

where TAIF is the Tail Area Influence Function defined in [4].

Replacing the normal model by the contaminated normal model $\Phi^\epsilon = (1 - \epsilon)\Phi_{\mu,\sigma} + \epsilon\delta_x$ and computing the derivative at $\epsilon = 0$ we obtain that

$$\begin{aligned} \text{TAIF}(x; t; \chi_g^2, \Phi_{\mu,\sigma}) &= \left. \frac{\partial}{\partial \epsilon} P_{\Phi^\epsilon}\{G_s(X_1, X_2) > t\} \right|_{\epsilon=0} \\ &= gP\{\chi_{g-1}^2 > t - (x - \mu)^2/\sigma^2\} - gP\{\chi_g^2 > t\} \end{aligned}$$

integrating now, we obtain the result. \square

Considering a scale contaminated normal (SCN) model

$$(1 - \epsilon)N(0, 1) + \epsilon N(0, \nu)$$

the VOM approximation is

$$p_g^F \simeq (1 - g\epsilon)P\{\chi_g^2 > t\} + g\epsilon \int_{-\infty}^{\infty} P\{\chi_{g-1}^2 > t - x^2\} d\Phi_{0,\nu}(x).$$

In Table 1 appear the *Exact* values (obtained through a simulation of 100.000 samples) and the VOM approximations when $\epsilon = 0'05$, $\nu = 2$ and $g = 3$.

Table 1 *Exact* and approximate p-values with $g = 3$

t	"exact"	approximate
6	0'149	0'148
8	0'077	0'076
10	0'042	0'042
12	0'024	0'025
14	0'016	0'016
16	0'011	0'011
18	0'007	0'008

To obtain the previous numerical results we had to deal with numerical integration. Sometimes, we would like to have analytic expressions of p_g^F to value the effect of contamination ϵ , etc. For this reason and for controlling

the relative error of the approximation, in the next section we shall compute the Saddlepoint approximation for the p-value of the Procrustes Statistic.

5.2 Saddlepoint Approximations for the p-value of the Procrustes Statistic

Using Lugannani and Rice formula, [8], for the sample mean of g independent square normal variables, we obtain the VOM+SAD approximation given in the next result.

Proposition 2. *Let $G_s(X_1, X_2)$ be the Procrustes statistic, that follows a χ_g^2 distribution when the underlying model is a normal distribution, $\Phi_{\mu, \sigma}$. If the null hypothesis H_0 holds, the saddlepoint approximation of the von Mises expansion, VOM+SAD approximation, for the functional tail probability (if F is close to the normal $\Phi_{\mu, \sigma}$) is*

$$P_F \{G_s(X_1, X_2) > t\} \simeq P \{\chi_g^2 > t\} - B + B \int_{-\infty}^{\infty} \frac{\sqrt{g}}{\sqrt{t}} e^{\frac{(t-g)(x-\mu)^2}{2t\sigma^2}} dF(x) \quad (2)$$

$$\text{where } B = \frac{g\sqrt{g}}{\sqrt{\pi}(t-g)} e^{-(t-g-g \cdot \log(t/g))/2}.$$

Proof. If $G_s(X_1, X_2)$ follows a χ_g^2 distribution, and Y_1, \dots, Y_g are g independent gamma distributions $\gamma(1/2, 1/2)$ with moment generating function M and cumulant generating function $K = \log M$, it is, following [8], [2] or [6],

$$\begin{aligned} P_{\Phi} \left\{ \frac{G_s(X_1, X_2)}{g} > t \right\} &= P \left\{ \frac{1}{g} \sum_{i=1}^g Y_i > t \right\} \\ &= 1 - \Phi_s(w) + \phi_s(w) \left\{ \frac{1}{r} - \frac{1}{w} + O(g^{-3/2}) \right\} \quad (3) \end{aligned}$$

where Φ_s and ϕ_s are the cumulative distribution and density functions of the standard normal distribution.

If K is the cumulant generating function, that is the functional of $\Phi_{\mu, \sigma}$,

$$K(\theta) = \log \int_{-\infty}^{\infty} e^{\theta(u-\mu)^2/\sigma^2} d\Phi_{\mu, \sigma}(u)$$

and z_0 is the (functional) saddlepoint, i.e., it is the solution of the equation $K'(z_0) = t$, the functionals that appear in (3) are

$$w = \text{sign}(z_0) \sqrt{2g \cdot (z_0 t - K(z_0))} = \sqrt{g} \text{sign}(z_0) \sqrt{2(z_0 t - K(z_0))} := \sqrt{g} w_1$$

$$r = z_0 \sqrt{g \cdot K''(z_0)} = \sqrt{g} z_0 \sqrt{K''(z_0)} := \sqrt{g} r_1.$$

As we saw before, the VOM approximation for the tail probability depends on the TAIF. To obtain the TAIF of $G_s(X_1, X_2)/g$ at $\Phi_{\mu, \sigma}$ we have to replace the model $\Phi_{\mu, \sigma}$ by the contaminated model $\Phi^\epsilon = (1 - \epsilon)\Phi_{\mu, \sigma} + \epsilon \delta_x$ in all the functionals in the right side of (3) that depend on $\Phi_{\mu, \sigma}$, and then to obtain the derivative at $\epsilon = 0$; this process is represented with a dot over the functional. Since $\phi'_s(w) = -\phi_s(w) w$ and $\phi_s(w) \leq 1$, we obtain that

$$\begin{aligned} \text{TAIF} \left(x; t; \frac{G_s(X_1, X_2)}{g}, \Phi_{\mu, \sigma} \right) &= \frac{\partial}{\partial \epsilon} P_{\Phi^\epsilon} \left\{ \frac{G_s(X_1, X_2)}{g} > t \right\} \Big|_{\epsilon=0} \\ &= -\phi_s(w) \dot{w} + \phi'_s(w) \dot{w} \left\{ \frac{1}{r} - \frac{1}{w} + O(g^{-3/2}) \right\} + \phi_s(w) \left\{ -\frac{\dot{r}}{r^2} + \frac{\dot{w}}{w^2} + O(g^{-3/2}) \right\} \\ &= \phi_s(w) \left[-\frac{w \dot{w}}{r} - \frac{\dot{r}}{r^2} + \frac{\dot{w}}{w^2} \right] + O(g^{-1}) \\ &= \phi_s(w) \left[-\frac{\sqrt{g} w_1 \sqrt{g} \dot{w}_1}{\sqrt{g} r_1} - \frac{\sqrt{g} \dot{r}_1}{g r_1^2} + \frac{\sqrt{g} \dot{w}_1}{g w_1^2} \right] + O(g^{-1}) \\ &= \frac{\phi_s(w)}{r_1} \left[-\sqrt{g} \cdot w_1 \dot{w}_1 \right] + O(g^{-1/2}) \end{aligned}$$

because the functionals w_1, \dot{w}_1, r_1 and \dot{r}_1 do not depend on g . Since

$$\dot{w}_1 = \text{sign}(z_0) \frac{2(\dot{z}_0 t - \dot{K}(z_0))}{2\sqrt{2(z_0 t - K(z_0))}} = \frac{\dot{z}_0 t - \dot{K}(z_0)}{w_1}$$

it will be

$$\text{TAIF} \left(x; t; \frac{G_s(X_1, X_2)}{g}, \Phi_{\mu, \sigma} \right) = \frac{\phi_s(w)}{r_1} \sqrt{g} \left[\dot{K}(z_0) - \dot{z}_0 t \right] + O(g^{-1/2}). \quad (4)$$

Hence, we have to compute the influence functions $\dot{K}(z_0)$ and \dot{z}_0 . To do this, because

$$K'(\theta) = \frac{\int_{-\infty}^{\infty} e^{\theta(u-\mu)^2/\sigma^2} \left(\frac{u-\mu}{\sigma} \right)^2 d\Phi_{\mu, \sigma}(u)}{\int_{-\infty}^{\infty} e^{\theta(u-\mu)^2/\sigma^2} d\Phi_{\mu, \sigma}(u)}$$

from the saddlepoint equation, $K'(z_0) = t$, we obtain

$$\int_{-\infty}^{\infty} e^{z_0(u-\mu)^2/\sigma^2} \left[\left(\frac{u-\mu}{\sigma} \right)^2 - t \right] d\Phi_{\mu,\sigma}(u) = 0.$$

Replacing again the model by the contaminated model $\Phi^\epsilon = (1-\epsilon)\Phi_{\mu,\sigma} + \epsilon\delta_x$ before obtaining the derivative at $\epsilon = 0$, and making the change of variable $(u-\mu)/\sigma = y$, we obtain

$$\dot{z}_0 \left[\int_{-\infty}^{\infty} e^{z_0 y^2} y^4 d\Phi_s(y) - t \int_{-\infty}^{\infty} e^{z_0 y^2} y^2 d\Phi_s(y) \right] + e^{z_0(x-\mu)^2/\sigma^2} \left[\left(\frac{x-\mu}{\sigma} \right)^2 - t \right] = 0$$

i.e.,

$$\dot{z}_0 = \frac{1}{2} t^{-5/2} e^{\frac{(t-1)(x-\mu)^2}{2t\sigma^2}} \left[t - \left(\frac{x-\mu}{\sigma} \right)^2 \right].$$

In a similar way, we obtain that

$$\dot{K}(z_0) = \frac{3}{2} t^{-1/2} e^{z_0(x-\mu)^2/\sigma^2} - \frac{1}{2} t^{-3/2} e^{z_0(x-\mu)^2/\sigma^2} \left(\frac{x-\mu}{\sigma} \right)^2 - 1.$$

Also it is

$$r_1 = z_0 \sqrt{K''(z_0)} = \frac{t-1}{\sqrt{2}} \quad \text{and} \quad \phi_s(w) = \frac{1}{\sqrt{2\pi}} e^{-g \cdot (t-1-\log t)/2}.$$

Therefore, from (4), it will be

$$\text{TAIF} \left(x; t; \frac{G_s(X_1, X_2)}{g}, \Phi_{\mu,\sigma} \right) = A \left(\frac{1}{\sqrt{t}} e^{\frac{(t-1)(x-\mu)^2}{2t\sigma^2}} - 1 \right) + O(g^{-1/2})$$

where

$$A = \frac{\sqrt{g}}{\sqrt{\pi}(t-1)} e^{-g \cdot (t-1-\log t)/2}.$$

From (1), we obtain now the VOM+SAD approximation for the p-value of the test statistic $G_s(X_1, X_2)/g$,

$$P_F \left\{ \frac{G_s(X_1, X_2)}{g} > t \right\} \simeq P \{ \chi_g^2 > g t \} - A + A \int_{-\infty}^{\infty} \frac{1}{\sqrt{t}} e^{\frac{(t-1)(x-\mu)^2}{2t\sigma^2}} dF(x)$$

and from this, we obtain the approximation (2) for the test statistic $G_s(X_1, X_2)$.
 \square

If F is the location contaminated normal mixture (LCN),

$$F = (1 - \epsilon) N(0, 1) + \epsilon N(\theta, 1)$$

the VOM+SAD approximation is

$$P_F \{G_s(X_1, X_2) > t\} \simeq P \{\chi_g^2 > t\} + \epsilon B \left[e^{-(1-t/g)\theta^2/2} - 1 \right].$$

In Table 2 appear the *Exact* values (obtained through simulation of 100.000 samples), the VOM and the VOM+SAD approximations when $\epsilon = 0'01$, $\theta = 1$ and $g = 5$.

Table 2 *Exact and approximate p-values with $g = 5$*

t	"exact"	VOM appr.	VOM+SAD appr.
9	0'1125	0'1129	0'1136
11	0'0538	0'0539	0'0545
13	0'0251	0'0249	0'0253
15	0'0114	0'0112	0'0115
17	0'0050	0'0049	0'0051
19	0'0022	0'0022	0'0023

Corollary 1. *To test the null hypothesis H_0 that there is no systematic difference between the standardized Configurations X_1 and X_2 with $p \times k$ landmarks (i.e., they belong to the same group) using the Procrustes statistic $G(X_1, X_2)$ and assuming that the error difference between Configurations*

$$\frac{X_2 - X_1}{\eta}$$

follow a scale contamination normal model $(1 - \epsilon)N(0, 1) + \epsilon N(0, \nu)$, the VOM+SAD approximation for the tail probability (p-value) is

$$P\{G_s(X_1, X_2) > t\} \approx P\{\chi_g^2 > t\} + \epsilon \frac{g^{3/2}}{\sqrt{\pi}(t - g)} \left[\frac{\sqrt{g}}{\sqrt{t - \nu^2(t - g)}} - 1 \right] \\ \cdot \exp \left\{ -\frac{1}{2} \left(t - g - g \cdot \log \frac{t}{g} \right) \right\} \quad (5)$$

where $g = kp - k(k + 1)/2 - 1$. It must be $p > (k + 1)/2 + 1/k$ and obviously an integer.

Then, if $k = 2$, it is $g = 2p - 4$ and $p > 2$. And if $k = 3$, it is $g = 3p - 7$ and $p \geq 3$.

6 Applications

In this section we are going to consider the following Example in which we make two comparisons using the previous theory.

Example 2. We are going to consider two test to check if there are or not significance differences between two arrows of *Notch tips and bay leaves*, of Solutrense period, arrows that were found in caves of Asturias (Spain). We shall make this analysis using a photo of the “Museo Arqueológico de Asturias” (Oviedo), included in QGIS as a raster layer, Fig. 16.

In this figure we see large differences among the arrows except in two pairs of arrows: Arrows 1 and 3 and arrows 5 and 6, Fig. 17. Hence, we are going to test the null hypothesis of no significance differences between arrows 1 and 3, and then, with another test, we shall check the null hypothesis of no significance differences between arrows 5 and 6.

To do this, we first create the polygons in QGIS marking the landmarks with the mouse. We consider $p = 7$ landmarks. Also, with QGIS we export the coordinates of the landmarks, that are:

Arrow 3

```
punta3<-matrix(c(
```

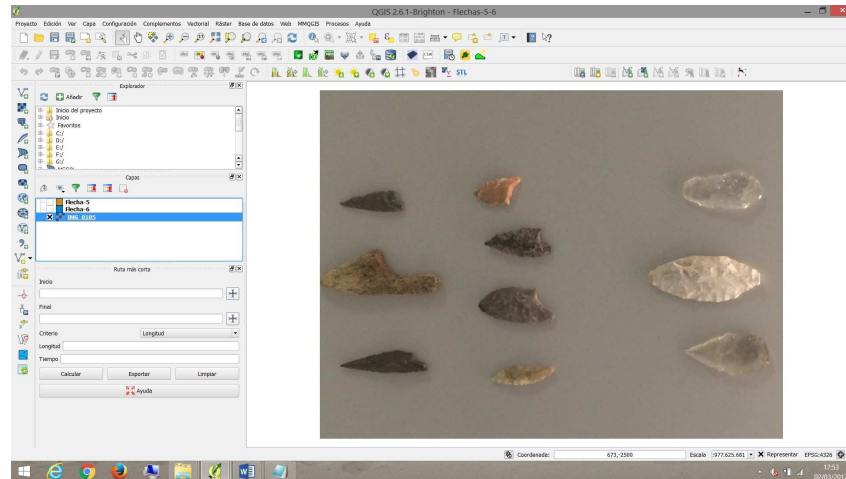


Fig. 16 Arrows in QGIS

```
151.77884,-794.21946,
384.34151,-714.48369,
533.84608,-706.17788,
543.81305,-756.01273,
587.00326,-794.21946,
583.68094,-842.39315,
384.34151,-849.03780),
ncol=2,byrow = T)
```

Arrow 1

```
punta1<-matrix(c(
157.59291,-1934.60710,
444.97392,-1841.58204,
640.99102,-1848.22668,
650.95799,-1891.41689,
734.01609,-1917.99548,
735.67725,-1966.16918,
428.36230,-1977.79731),
ncol=2,byrow = T)
```

Arrow 6

```
punta6<-matrix(c(
1170.17428,-1072.54821,
1423.18465,-971.34406,
1550.56229,-974.83386,
1557.54188,-1039.39513,
1606.39906,-1074.29311,
1608.14396,-1156.30337,
1410.97036,-1170.26256),
```

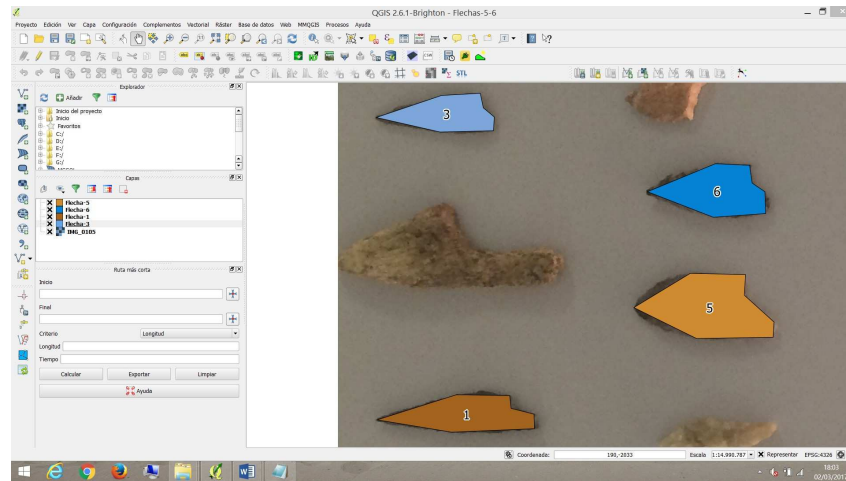


Fig. 17 Arrows as polygons in QGIS


```
ncol=2,byrow = T)

Arrow 5

punta5<-matrix(c(
1119.57220,-1510.51789,
1327.21520,-1383.14025,
1533.11329,-1386.63005,
1529.62350,-1465.15051,
1639.55214,-1564.60976,
1637.80724,-1618.70164,
1349.89889,-1625.68123),
ncol=2,byrow = T)
```

Comparison between Arrows 1 and 3

After removing the effect of **Size** (scale), **Location** (translation) and **Orientation** (rotation) to standardize the individuals, we match them at the common centroid obtaining the polygons of Fig. 18.

The minimum *Residual Distance* between configurations (arrows), i.e., the value of the *Procrustes statistic* for testing the null that “No significance differences exist between arrows 1 and 3” is 0.01567681:

```
> tamapunta1<-sqrt(sum(apply(punta1,2,var)*(7-1)))
> spunta1<-scale(punta1/tamapunta1,scale=F)
> tamapunta3<-sqrt(sum(apply(punta3,2,var)*(7-1)))
> spunta3<-scale(punta3/tamapunta3,scale=F)
> library(shapes)
```

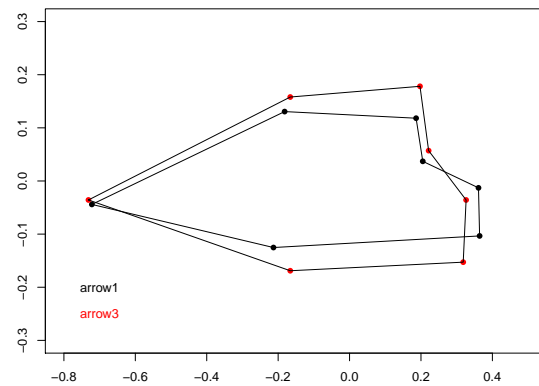


Fig. 18 Polygons of arrows 1 and 3

```

> (procdist(spunta1,spunta3,type="partial"))^2
[1] 0.01567681

> sd(spunta1-spunta3)
[1] 0.03502222

```

(1)

Because of (1), choosing $\eta = 0.03502222$, we shall obtain a standard normal distribution for $(X_2 - X_1)/\eta$ and hence

$$G(X_1, X_2) \approx \eta^2 \chi_g^2$$

being $g = 2p - 4 = 10$. Then, the p-value of this classical test will be

$$\begin{aligned}
 P(\text{Procrus.Stat.} > 0.01567681) &= P(\chi_{10}^2 > 0.01567681/(0.03502222^2)) \\
 &= P(\chi_{10}^2 > 12.78116) = 1 - pchisq(12.78116, 10) = 0.2361661
 \end{aligned}$$

accepting the null hypothesis of no significance differences between Arrows 1 and 3.

Nevertheless, using the Mahalanobis distance we can conclude that the errors do not follow a multivariate normal distribution,

```

> dipuntas2<-mahalanobis(spunta1-spunta3,colMeans(spunta1-spunta3),
+ var(spunta1-spunta3))
> ks.test(dipuntas2,"pchisq",7)

One-sample Kolmogorov-Smirnov test

data: dipuntas2
D = 0.76612, p-value = 8.265e-05
alternative hypothesis: two-sided

```

Hence, to assume a common η for all the c_{ij} such that

$$\frac{X_2 - X_1}{\eta} \rightsquigarrow N(0, 1)$$

is unrealistic. It is better to consider a model

$$0'9N(0, 1) + 0'1N(0, \nu)$$

and to use the VOM+SAD approximation (5), programmed into the R function `apro3(g, ν , ϵ , t)` to compute the p-value. Obtaining from the data $\eta = 0.04020902$ and $\nu = 0.032261$, we have

$$P(\text{Procrus.Stat.} > 0.01567681) = P(\text{Procrus.Stat.}/(0.04020902^2) > 0.01567681/(0.04020902^2))$$

$$= P(Procrus.Stat./(0.04020902^2) > 22.95901) = P(G_s(X_1, X_2) > 22.95901) = 0.006318776$$

because

```
> apro3(10,0.032261,0.1,22.95901)
[1] 0.006318776
```

Then, because

$$(VOM + SAD) \text{ p-value} = 0.006318776$$

we reject, in a more robust way, the null hypothesis of no significance differences between Arrows 1 and 3.

Comparison between Arrows 5 and 6

After removing the effect of **Size** (scale), **Location** (translation) and **Orientation** (rotation) to standardize the individuals we match them at the common centroid having the polygons of Fig. 19.

The minimum *Residual Distance* between configurations (arrows), i.e., the value of the *Procrustes statistic* for testing the null that “No significance differences exist between arrows 5 and 6” is 0.03711933,

```
> tamapunta5<-sqrt(sum(apply(punta5,2,var)*(7-1)))
> spunta5<-scale(punta5/tamapunta5,scale=F)
```

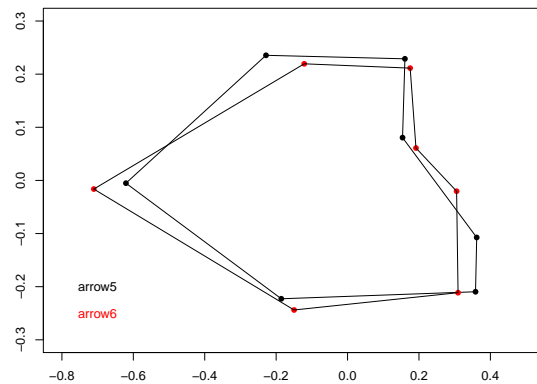


Fig. 19 Polygons of arrows 5 and 6

```

> tamapunta6<-sqrt(sum(apply(punta6,2,var)*(7-1)))
> spunta6<-scale(punta6/tamapunta6,scale=F)
> library(shapes)
> (procdist(spunta5,spunta6,type="partial"))^2
[1] 0.03711933

> sd(spunta5-spunta6)
[1] 0.05343598

```

(2)

Because of (2), if $\eta = 0.05343598$ we shall obtain a standard normal distribution for $(X_2 - X_1)/\eta$. Then,

$$G(X_1, X_2) \approx \eta^2 \chi_g^2.$$

and hence, the p-value of this classical test will be

$$P(\text{Procrus.Stat.} > 0.03711933) = P(\chi_{10}^2 > 0.03711933/(0.05343598^2))$$

$$= P(\chi_{10}^2 > 12.99968) = 1 - pchisq(12.99968, 10) = 0.2236897$$

accepting the null hypothesis of no significance differences between Arrows 5 and 6.

Nevertheless, using the Mahalanobis distance we can conclude that the errors do not follow a multivariate normal distribution,

```

> dipuntas<-mahalanobis(spunta6-spunta5,colMeans(spunta6-spunta5),
+ var(spunta6-spunta5))
> ks.test(dipuntas,"pchisq",7)

```

One-sample Kolmogorov-Smirnov test

```

data: dipuntas
D = 0.76677, p-value = 8.093e-05
alternative hypothesis: two-sided

```

Then, to assume a common η for all the c_{ij} such that

$$\frac{X_2 - X_1}{\eta} \rightsquigarrow N(0, 1)$$

is unrealistic. It is better to consider a model

$$0'9N(0, 1) + 0'1N(0, \nu)$$

and to use the VOM+SAD approximation (5), programmed as the R function `apro3(g, ν , ϵ , t)` to compute p-values. From the data we obtain $\eta = 0.06834322$ and $\nu = 0.0389347$ and hence,

$$P(Procrus.Stat. > 0.03711933) = P(Procrus.Stat./(0.06834322^2) > 0.03711933/(0.06834322^2))$$

$$= P(G_s(X_1, X_2) > 7.947111) = 0.5405565$$

because

```
> apro3(10,0.0389347,0.1,7.947111)
[1] 0.5405565
```

Then, because

$$(VOM + SAD) \text{ p-value} = 0.5405565$$

we finally accept, in a more robust way, the null hypothesis of no significance differences between Arrows 5 and 6.

7 Conclusions

Classical Morphometric Analysis based on Landmarks is reviewed from a descriptive and inferential point of view. Because both are based on sample means and least-squares they are not robust.

We first robustify the descriptive measures proposing robust ones. Then we consider a Contaminated Normal Model distribution instead of a classical Normal one to make robust inferences. Namely, for this mixture model we obtain an von Mises approximation for the p-value of a test for the null hypothesis of no significance differences between two individuals based on their shapes.

We also obtain a very accurate saddlepoint approximation of this von Mises approximation. We conclude the paper with some applications using QGIS as Geographical Information System.

Acknowledgements This work is partially supported by Grant HAR2015-68876-P from Ministerio de Economía y Competitividad (Spain).

References

1. Claude J (2008) Morphometrics with R. Springer
2. Daniels HE (1983) Saddlepoint approximations for estimating equations. *Biometrika* 70:89–96

3. Dryden IL, Mardia KV (2016) *Statistical Shape Analysis with Applications in R*. John Wiley and Sons
4. Field CA, Ronchetti E (1985) A tail area influence function and its application to testing. *Commun Stat* 4:19–41
5. García-Pérez A (2006) Chi-square tests under models close to the normal distribution. *Metrika* 63:343–354
6. Jensen JL (1995) *Saddlepoint approximations*. Clarendon Press, Oxford
7. Langron SP, Collins AJ (1985) Perturbation theory for Procrustes Analysis. *J R Stat Soc Ser B Stat Methodol* 47:277–284
8. Lugannani R, Rice S (1980) Saddle point approximation for the distribution of the sum of independent random variables. *Adv Appl Probab* 12:475–490
9. O’Higgins P, Dryden IL (1993) Sexual dimorphism in hominoids: further studies of craniofacial shape differences in Pan, Gorilla, Pongo. *J Hum Evol* 24:183–205
10. R Development Core Team (2016), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Viena, Austria. URL <http://www.R-project.org>
11. Sibson R (1979) Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *J R Stat Soc Ser B Stat Methodol* 41:217–229